Information Retrieval Evaluation in Knowledge Acquisition Tasks

Yasin Ghafourian[1,2], Petr Knoth[1,3] and Allan Hanbury[2]

[1] *Research Studios Austria Forschungsgesellschaft (RSA FG), Vienna, Austria*
[2] *Vienna University of Technology (TU Wien), Vienna, Austria*
[3] *The Open University, Milton Keynes, The United Kingdom*

Abstract

The Cranfield Paradigm is a widely adopted and the de-facto standard approach to the evaluation of IR systems. However, this approach does not inherently support situations in which the user is acquiring knowledge (is learning) during an information seeking session consisting of the submission of a sequence of queries into an information retrieval system. More specifically, during a situation in which the retrieval of a particular document at the beginning of a session can be considered not relevant (due to the user's lack of knowledge), while it can be considered relevant at a later point in the session (once the user acquired all required prerequisite knowledge). In this position paper, we reflect on the limitations of the Cranfield Paradigm in the context of *knowledge acquisition tasks* and propose several alternatives. These alternatives are based on the notion of evaluating a session consisting of a sequence of individual queries created to address a specific information need as part of a knowledge acquisition task.

Keywords

Information Retrieval, Knowledge Delta, Personalisation, Evaluation, Knowledge Acquisition, Search as Learning,

1. Personalisation with Knowledge

There are a variety of reasons for users to interact with information retrieval (IR) systems and multiple classification of information seeking behaviour, for instance [1]. In this paper, we focus on *knowledge acquisition tasks*, which we define as tasks in which the user aims to acquire knowledge as part of a sequence of interactions with the IR system about a subject or concept they are not yet fully familiar with. In knowledge acquisition tasks, the interaction of the user with the IR system should be determined by both the interest of the user as well as their prior knowledge. Users with different background knowledge of the topic in question might benefit from different paths through the resources indexed by the IR system. Additionally, their ability to compose and formulate queries on a topic they are not fully familiar with might be limited. In

WEPIR 2021: The Third Workshop on Evaluation of Personalisation in Information Retrieval at CHIIR 2021, 19 March 2021. Virtual event.

yasin.ghafourian@researchstudio.at (Y. Ghafourian); petr.knoth@researchstudio.at (P. Knoth); hanbury@ifs.tuwien.ac.at (A. Hanbury)
http://www.open.ac.uk/people/pk3295 (P. Knoth); https://informatics.tuwien.ac.at/people/allan-hanbury (A. Hanbury)
0000-0001-9683-9748 (Y. Ghafourian); 0000-0003-1161-7359 (P. Knoth); 0000-0002-7149-5843 (A. Hanbury)

CEUR Workshop Proceedings (CEUR-WS.org)

knowledge acquisition tasks, users are overcoming their *knowledge deltas* as part of a sequence of information retrieval tasks.

In an attempt to measure and predict the user's background knowledge (referred to as the knowledge state) and knowledge gain resulting from a search process, Yu et al. [2] employ supervised classification methods based on a variety of user interaction features extracted from search logs. The ground truth, which consists of users' background knowledge before and after search sessions for a predefined set of topics, is established by means of the formulated knowledge tests conducted by the users. Each user is assigned their knowledge state and knowledge gain as *low,moderate* or *high.* This constitutes one example of how a knowledge gain can be measured. However, achieving a knowledge gain does not necessarily mean the bridging of a knowledge delta.

In a keynote speech on information search processes Vakkari [3] discussed how users go through a three-stage process of *restructuring*, *tuning* and *assimilation*, over which the users converge from typically vague search terms to more specific search terms over complex search tasks that involves learning. One could see this a theoretical framework for for overcoming a knowledge delta.

To illustrate how the presence of users with different background knowledge in a knowledge acquisition task will drive the search session to different paths, we consider the following scenario as a straightforward example from the domain of *Natural language processing (NLP).* Let's assume that two users, namely *user 1* and *user 2* both want to learn about *"Latent Semantic Analysis" (LSA). user 1* is more experienced and has some prior knowledge in the field of *"Information Retrieval"* and *"Linear Algebra"* while on the contrary, *user 2* is new to the field and carries only some prior knowledge of *"Probability Theory".* Consequently, as the goal of the knowledge acquisition task is to gain knowledge of *"Latent Semantic Analysis"*, *user 1*'s information need can be satisfied by a session comprising of only one or two queries and reading some documents about the main subject. This is in contrast to *user 2* who realises that he lacks some prerequisite knowledge that is required to understand the desired topic. This user then reformulates his/her query several times so as to first learn with the goal to first understand the necessary background information on *"Information Retrieval"* and *"Linear Algebra".*

Although the discussed example is also subject to arguments with regard to, for example, the extent to which *user 1* is familiar with the background matters that also affects the knowledge acquisition task, the main issue here is that within the Cranfield Paradigm, IR systems are evaluated based on the assumption that a user's information need is represented well in his query and it's not taken into account that users are unable to express queries about topics they are not familiar with and will not properly formulate their information need (i.e. they don't know what exists in the extensive world wide web) [4]. This problem calls for going beyond only considering the query as the information carrier. In addition, it's noteworthy to mention that in assessing the utility of a document, users are influenced by many factors that go beyond merely topical relevance [5], which can depend on search, document or user context as well. Liu et al. [6] provides a comprehensive list of the contexts used in both contextual and personalised IR systems as well as the research done to investigate the effect of users' subject domain knowledge and task topic knowledge as user contexts on their search behaviour.

In our scenario, the user's context of interest will also be the user's domain knowledge, that also influences the user's perception of relevance of the presented document(s). Referring to our scenario, in response to a query to learn about *"Latent Semantic Analysis"*, documents that are discussing topics such as *matrix product* and *Singular Value Decomposition (SVD)* are deemed irrelevant by *user 1* who already is in possession of the required background knowledge to learn about *LSA*. However, the same documents could be in favour of *user 2* with the same submitted query who only has some basic knowledge of the field. What happens here is that disregarding a user's prior knowledge will present the user with a ranked list of documents that might be well matching the user's query in terms of topical relevance but do not necessarily reflect the user's information need. For these users, the answer is not the lookup of a single relevant document, but rather a suitable ordered sequence of relevant documents specific to their knowledge level.

IR systems that effectively utilise user related context to provide user specific ranked results are called personalised information retrieval (PIR) systems. A system incorporating a user's knowledge level to provide relevant lists of documents to a query at each search iteration of a knowledge acquisition search task is also a personalised system. However, this system is different than personalisation systems that infer the user's interests and incorporate those in search queries or re-rank the results incorporating those interests to make the ranked list closer to the user's interests [7]. In our envisioned evaluation framework, a user's knowledge is not used directly by the system to retrieve documents that match what the user already knows. On the contrary, it is used as a foundation to retrieve the documents that will be most relevant to the user to bridge their knowledge delta.

As mentioned earlier, in evaluation of an information retrieval system that supports the search task through personalisation, solely measuring the topical relevance is not sufficient. Effectively evaluating the performance of the IR system in our scenario of interest requires measuring the relevance of the ranked list of documents to the user's submitted query with respect to the knowledge level of the user in the target domain (which for this paper we assume is already calculated and stored in a user profile within the system as that discussion falls in the scope of different stages in the implementation of personalisation [7] and is beyond the scope of this paper). The difference between evaluation of this system with other personalised systems that incorporate user interests is that in those personalisation systems, one can treat every topic/query independently from the perspective of the evaluation. However, the order in which the user submits queries and interacts with resources matters and determines the performance of the system.

In the rest of this paper we first address in section 2 the shortcomings of current performance measurement frameworks for the described use case. In section 3, we propose some prospective solutions to build an evaluation framework for the evaluation of a system that incorporates user's knowledge in a knowledge acquisition task.

2. Why are Existing Solutions not Sufficient?

An information retrieval system's (IRS) evaluation is done with the objective of gauging it's global performance. The traditional evaluation of IR systems is based on the laboratory model initiated by Cleverdon [8]. The series of projects and retrieval experiments by Cleverdon et al. in controlled laboratory-like settings is the paradigm evaluation of system-oriented IR research and provided the foundation for the evaluation of IR systems.

The Cranfield Paradigm of IR evaluation is based on a test collection which is mainly composed of three components. The components are 1) a static set of documents, 2) a set of information needs/topics and 3) a relevance judgment file indicating the relevance of each test document with regard to each topic. The main evaluation measures in the Cranfield Paradigm are *precision* and *recall* based on which several other measures are also used in TREC [9] such as *Mean Average Precision, Precision@X, ...* and other measures introduced along the research for IR evaluation within the Cranfield Paradigm.

One aspect that makes the laboratory-based evaluation models of an IR system – which typically measure the performance of the system in terms of *MAP*, *F-measure*, *NDCG*, etc – not ideal for the evaluation of a personalised IR system is that the user aspects and the surrounding contexts are not considered within the evaluation framework. As a case in point, we can refer back to our scenario where user is trying to acquire new information in a domain where he/she has very little prior knowledge about it. This brings the generalisability of a laboratory-based evaluation model under question as a real-life IR task is a social activity to which a user's subjectivity and cognitive aspects are inherent [10, 11].

Another aspect of the existing IR evaluation frameworks that makes them unsuitable for evaluating exemplified knowledge acquiring situations is that IR systems are evaluated in a way in which the response of the system to each query is treated independently. However, this simplification is not suitable for situations in which the user is acquiring knowledge throughout the process of searching and investigating the retrieved documents. During a search session, a document that might not be relevant for a user due to the lack of his/her knowledge at the beginning of the session might actually become relevant at a later stage as he/she is expanding his/her knowledge level in the field during the session.

The aforementioned problems have resulted for the evaluation of PIR systems to remain mostly in the user-centered fashion, which relies on user studies and involvement of real users in the experiments. Besides the advantage of the direct assessment of the subjective and cognitive aspects of an IRS, user-centered evaluation models have the serious drawback of not being reproducible and not providing the possibility of comparative study across several personalised IR systems.

According to Pasi [12], the evaluation of PIR systems can be explored from two aspects. First is the user profile aspect with a focus on the assessment of the quality of the user profiles, and second is the evaluation of the PIR system's effectiveness. There is still research needed to reach an agreed upon framework for the evaluation of PIR systems [6]. However, Tamine et al. [11] outline three main approaches to set up an evaluation setting for personalised systems, which we are also considering to propose as an evaluation framework for our knowledge acquisition task considering a user's knowledge. The first approach is the extended laboratory-based approach to account for the presence of contextual factors. The second approach is a simulation-based approach in which studies simulate users and interactions by means of retrieval scenarios. Finally, the third approach evaluates the system's performance qualitatively incorporating real users who undertake the qualitative evaluation. The evaluation methods of this approach include interviews, forms and observations of user's behavior with respect to a retrieval task.

Adopting the approach of extended laboratory-based evaluation of PIR systems, and with the aim of developing an evaluation methodology with a corresponding publicly available test collection that enables the possibility of repeatable evaluation of personalised search systems, recently Pasi et al. [13] have introduced an experimental framework for the creation of test collections to facilitate repeatable laboratory-based evaluation of PIR systems. Their procedure is also described in more detail later in [14]. The test collection, which was originally developed in [13], was later also released to participants of the personalised information retrieval Lab at CLEF 2018 and 2019 (PIR-CLEF 2018, PIR-CLEF 2019) [15, 16]. These Labs were commenced to provide an initiative aimed at both providing and critically analysing new approaches to the evaluation of Personalisation in IR. However, the problem with the dataset provided in [13] for the scenario at hand is that the user profiles in this test collection are made using all the documents that the users have marked relevant during their search session [15], and thus are made under the assumption that all documents remain relevant to the user's topic until the end of the session. This is in contrast to what was discussed earlier: during a knowledge acquisition search task, a document might not be relevant at first stages of the session and become relevant later or vice versa. Besides, Bai et al. [17] report that the aforementioned test collection comprises not the entire set of user actions in the search logs, thereby making it challenging to analyze user's preferences by their actions. In our case, this will make it difficult to assess the approximate acquired knowledge of a user after every submitted query (or more generally, after ever iteration) during the session.

The key research question here is: how to design an evaluation framework that would be capable of assigning a session, which consists of a sequence of queries and consumed documents (iterations) on a given topic, a score, such that IR systems helping the user to more quickly overcome their knowledge delta would receive a higher one. A subsequent question is: can we develop a ground truth test collection for a personalised knowledge acquisition system to be evaluated on and thereby provide the opportunity of fair laboratory-based comparison across such systems through a reusable dataset.

3. Prospective Solutions

There are multiple options that might provide a more suitable environment for evaluating knowledge acquisition tasks. Specifically, we explore and propose three options: 1) Online evaluation approach, 2) Prerequisite-labeled relevance judgements approach and 3) Session-based evaluation approach.

Online evaluation approach: One option might be to measure the Click-through rate (CTR) or to use similar online evaluation metrics. Under this option, the algorithm that maximises the CTR across sessions is performing the best. However, this approach (due to its online nature) does not allow the testing of large numbers of parameters through parameter tuning in the same way as it is possible for experiments conducted offline with a ground truth. This also accounts for the irreproducibility of the results with the online evaluation. Furthermore, click through data has also been critiqued in terms of its proposed associations with user relevance assessment [18].

Prerequisite-labeled relevance judgements approach: Another option might be to follow the Cranfield Paradigm and provide a framework in which each relevance judgement is annotated with the prerequisite knowledge required for this document to be considered a hit. A response of the IR system to each query would be the function of both the query itself and the user's knowledge, expressed in the form of the knowledge prerequisites that the user possesses. The disadvantage of this approach is that it might be rather resource intensive to produce such prerequisite-labeled relevance judgements when compiling the ground truth. An example of this approach could be the case where the ground truth data might contain information such that the resource describing *"Latent Semantic Analysis"* is only considered relevant provided that the user has already heard about *"Information Retrieval"*. Essentially, it is prerequisite dependent). This method could work well provided that the annotators are able to understand quickly (or with some support) the prerequisites for any given retrieved resource/document with regards to a query when they are composing the GT.

As an slight adaption to the relevance judgments for the GT for the discussed laboratory-based evaluation method, and considering that annotators have access to prerequisites for the given retrieved document to be annotated, one might want to consider the application of graded relevance judgments [19] rather than the traditional use of a binary scale. This is in recognition of our observation that it might be hard to classify documents that might be too complex for a particular user to fully understand yet still useful using a binary schema. To shed more light on this, lets assume a case where a user with only basic knowledge of *"mathematics"* is searching to learn about *"Page Rank"*. Some of the different prerequisites of the target topic – in this case *"Page Rank"* – to learn about prior to learning the target topic so as to understand the target topic better are *"Graph Theory"* and *"Information Retrieval"*. Accordingly, when the user begins the search session, considering his/her initial knowledge level, having a multi-label relevance judgment system with for example 3 levels (0:irrelevant, 1:semi-relevant, 2:relevant) allows for the documents about the main topic to be labeled as 1 (semi-relevant) compared to the documents that describe the prerequisite subjects *"Graph Theory"* and *"Information Retrieval"* that are more matching to the information need of the user at the beginning of his/her search session and can be labeled as 2 (relevant). As the session moves forward, and the user attains knowledge about prerequisite topics, the relevance label of prerequisite documents within *"Graph Theory"* and *"Information Retrieval"* with regard to the knowledge of the user could be changed to 1 (semi-relevant) at further iterations while at the same time, the relevance of the documents closely related to the main subject increase and they can be labeled with 2 (relevant). Having a graded relevance judgment system allows us to separate documents within domains that are completely irrelevant to the target topic and it's prerequisites from the documents that are relevant but their relevance grades can change during the session.

Session-based evaluation approach: Another way to look at an evaluation of a knowledge acquisition information seeking task is to look for the shortest sequence of queries and accessed documents that will lead the user to the state in which the user overcomes the knowledge delta and becomes familiar with the target topic. Such evaluation strategy follows the idea of minimising the number of queries and documents the user needs to visit in order to obtain sufficient knowledge about the topic.

To illustrate, let's suppose the information needed by the user is to understand *"Page Rank"*, but the user is unaccustomed to the concept of *"Directed Graphs (DGs)"*. The user can use an IR system **A** which retrieves first a document that defines *"Page Rank"*. The user will then realize that she cannot easily proceed with comprehending the document at hand and that reading this page calls for some background knowledge of *"Graph Theory"* and *"Directed Graphs"* , and so the user issues another query. A better IR system **B** (which uses personalisation support through exploiting user's knowledge) will retrieve a document explaining *"Directed Graphs (DGs)"* first and then a document about *"Page Rank"* to the same user. System **B** should be scored higher, because it satisfied the information needed (as opposed to merely satisfying the topical relevance) in a shorter session (1 query, 2 documents). On the contrary, to another user with sufficient knowledge from the domain of *"Graph Theory"*, documents describing it in the ranked list in response to the query being *"Page Rank"* are considered irrelevant. One way to think of an evaluation measure to develop within this framework is to set up a formula that aggregates the costs associated with the performance of each system. As a case in point, there are three costs associated with running each iteration of a session. These are: the cost of formulating and running the query, the cost of scanning through the list of documents and finally the cost of consuming a seemingly relevant document. These costs could be specified in terms of suitable quantities and be combined along with other costs in a formula that measures the performance of an approach in terms of it's costs.

4. Conclusion

This paper argues that there is a need for IR evaluation metrics that would be suitable for knowledge acquisition tasks which use personalisation support through exploiting user's knowledge. These are tasks in which a user aims to overcome their knowledge delta as part of a sequence of interactions with an IR system. We elaborated on three directions in which it might be possible to address this problem and expressed their limitations. In addition to the fairly standard online approach, the prerequisite-labeled relevance judgements and session-based approaches constitute feasible directions for future work in the area of evaluating knowledge acquisition information seeking tasks.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860721.

References

[1] Y. Li, N. J. Belkin, A faceted approach to conceptualizing tasks in information seeking, Information processing & management 44 (2008) 1822–1837.

[2] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, S. Dietze, Predicting user knowledge gain in informational search sessions, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 75–84.

[3] P. Vakkari, Information search processes in complex tasks, in: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, 2018, pp. 1–1.

[4] D. Kelly, N. J. Belkin, A user modeling system for personalized interaction and tailored retrieval in interactive ir, Proceedings of the American Society for Information Science and Technology 39 (2002) 316–325.

[5] G. Pasi, Contextual search: issues and challenges, in: Symposium of the Austrian HCI and Usability Engineering Group, Springer, 2011, pp. 23–30.

[6] J. Liu, C. Liu, N. J. Belkin, Personalization in text information retrieval: A survey, Journal of the Association for Information Science and Technology 71 (2020) 349–369.

[7] M. R. Ghorab, D. Zhou, A. O'connor, V. Wade, Personalised information retrieval: survey and classification, User Modeling and User-Adapted Interaction 23 (2013) 381–443.

[8] C. Cleverdon, The cranfield tests on index language devices, in: Aslib proceedings, MCB UP Ltd, 1967.

[9] D. Harman, Overview of the fourth text retrieval conference (trec-4), NIST Special Publication (1996) 1–23.

[10] P. Borlund, The iir evaluation model: a framework for evaluation of interactive information retrieval systems, Information research 8 (2003) 8–3.

[11] L. Tamine-Lechani, M. Boughanem, M. Daoud, Evaluation of contextual information retrieval effectiveness: overview of issues and research, Knowledge and Information Systems 24 (2010) 1–34.

[12] G. Pasi, Issues in personalizing information retrieval., IEEE Intell. Informatics Bull. 11 (2010) 3–7.

[13] G. Pasi, G. J. Jones, S. Marrara, C. Sanvitto, D. Ganguly, P. Sen, Overview of the clef 2017 personalised information retrieval pilot lab (pir-clef 2017), in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017, pp. 338–345.

[14] G. J. Jones, G. Pasi, A. Angiolillo, C. Sanvitto, A proposed method for laboratory-based evaluation of personalised information retrieval, Proceedings of WEPIR 2018 (2018).

[15] G. Pasi, G. J. Jones, K. Curtis, S. Marrara, C. Sanvitto, D. Ganguly, P. Sen, Overview of the clef 2018 personalised information retrieval lab (pir-clef 2018) (2018).

[16] G. Pasi, G. J. Jones, L. Goeuriot, L. Kelly, S. Marrara, C. Sanvitto, Overview of the clef 2019 personalised information retrieval lab (pir-clef 2019), in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 417–424.

[17] Q. Bai, Q. Zhang, Q. Hu, L. He, Ecnu at clef pir 2018: Evaluation of personalized information retrieval., in: CLEF (Working Notes), 2018.

[18] Q. Guo, E. Agichtein, Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior, in: Proceedings of the 21st international conference on World Wide Web, 2012, pp. 569–578.

[19] T. Sakai, On the reliability of information retrieval metrics based on graded relevance, Information processing & management 43 (2007) 531–548.